Visual Analytics of The Lord of the Rings Trilogy

Stefan Diener - 210008139

Abstract— This paper employed a visual approach to analyze the combined texts of the Lord of The Rings trilogy. The aim of this study was to show how using computational and visual methods in combination with human reasoning can lead to a deeper understanding of the source material. The results show that the visualization of sentiment analysis and character occurrences throughout the story are able to identify patterns and structure in the story. Moreover, word clouds have been shown to be a useful tool for text summarization and to deeper investigate the concept of good and evil in the story. Finally, network analysis suggested Frodo to be the story's protagonist and revealed further nuances about the relationships among the other main characters. However, many of the analyses are based on keyword counts, which have limitations mainly related to their inability of taking the context that the keyword appears in into account.

1 Problem Statement

The shift of modern communication to the digital domain is the catalyst for an ever-increasing volume of digital texts. Computational methods for visualizing and analyzing text data that are able to help humans understand these large text corpora are therefore becoming increasingly important in the scientific literature [1].

This study uses the famous Lord of the Rings (LOTR) novels by J. R. R. Tolkien, to show how the application of text processing techniques can lead to a deeper understanding of the text at hand. More precisely, it aims to answer the following research questions:

1. Does the visualization sentiment and character occurrences across time detect key parts and tuning points of the story?
2. How can text summarization with word clouds help identify and understand certain concepts in the text?
3. Can network analysis help detect the main characters in the story and reveal insights about their relationships to each other?

The data used for the following analysis is composed of two parts: first, the books of the LOTR trilogy in text format, and second, additional data on all characters. This data is suitable for answering the research questions because it covers the entire LOTR story because it is in an unstructured text format and because it contains a comprehensive list of all characters, which makes it possible to find the most important characters and analyze their relationships without prior knowledge of the story.

2 State of the Art

This section provides an overview of three papers that leverage computational analysis in tandem with effective visualizations to provide insight into the text corpus at hand.

In the first study, Mohammad [3] shows how sentiment analysis can be used to quantify and track the polarity and emotions of mail and books. Here, polarity refers to the positivity positive or negativity of a text, whereas emotions refer to the eight distinct measures: joy, sadness, anger, fear, trust, disgust, surprise, anticipation. To quantify each text along any of those dimensions, a lexicon-based approach is used, where each word of a text is compared to a dictionary that associates a vocabulary of words with a certain polarity or emotion. With this approach, he compares the polarity and emotional content of different kinds of mail, like love letters, hate mail, suicide notes, or emails written by men versus women.

One of the reasons why Mohammad's paper is highly relevant to the subsequent analysis of this study is because he not only applies these techniques to short texts like mail but also shows how these techniques can be applied to entire books. By calculating the emotions for each line in a novel, Mohammad is able to track the development of certain emotions across the story. Furthermore, uses a word cloud to visualize the words that are associated with a certain emotion, which helps to convey a deeper understanding of what these emotions are based on in the context of the story.

In the second paper, Heimerl, Lohmann, Lange, and Ertl [4] build a system called Word Cloud Explorer that aims to improve the power of basic word clouds for text analysis, by leveraging interactivity, natural language processing techniques, and context information, like e.g., part of speech tagging. To test their hypothesis, they conduct a qualitative user study, where a group of analytics professionals completed analytics tasks on three different corpora using the new software. They concluded that its main advantages were increased flexibility and intuitiveness. Hence, this study shows that combining word clouds with additional text processing techniques can be an effective way for exploring and analyzing text data.

In the third paper, Rydberg-Cox [5] analyzes a corpus of Greek tragedies with social network graphs. His goal is to discover quantifiable patterns about the tragedies and utilize visualizations of these networks to communicate the patterns. By representing each character as a node and connections between characters as edges in the network graph, he was able to identify four distinct patterns across the corpus.

This work shows how network analysis can be used to model relationships between characters, however, Rydberg-Cox also states that the limited number of characters in each tragedy strongly contributed to the emergence of clear patterns. Thus, when applying network analysis, it seems to be useful to think about whether the number of characters

included in the analysis can be limited in a meaningful way in order to increase the interpretability of the network graph.

3 Properties of the Data

The data of two main components: first, the books of the LOTR trilogy, and second, demographic data on all characters in the LOTR universe. The data was downloaded from a public GitHub repo that used https://archive.org and www.ageofthering.com to scrape this information from the web [2].

Each part of the LOTR trilogy is stored in a text file that represents the entire physical equivalent, from the title and contents to the footnotes at the very end. The structure of the LOTR series is somewhat unconventional and is as follows: First, the trilogy is divided into three parts "The Fellowship of the Ring", "The Two Towers", and "The Return of the King". Secondly, each part is divided into two so-called books, which adds up to a total of six books in the complete trilogy. Lastly, each of the six books contains between 9 and 12 chapters. In total, the LOTR series contains about 470 thousand words. And uses a vocabulary of ca. 12 thousand words.

One of the biggest challenges with this data is that it is inherently unstructured, with each of the three text files essentially containing only a very long string of characters. In conclusion, for further analysis, it is important to represent the data in a more structured way, that allows for the calculation of metrics e.g., keyword occurrences over time or per chapter.

By manually skimming through the text data, it seems like the data is a very accurate representation of the original books. Encoding errors, like missing, swapped, or double characters, additional white spaces within a word, or missing white spaces between words, seem to be very rare or potentially non-existent. However, typical for text data, it contains punctuation, capitalization, and many stop words, which are often not useful for analysis. This is exemplified by Figure 1, which shows that the top 20 words in the data are exclusively stop words.

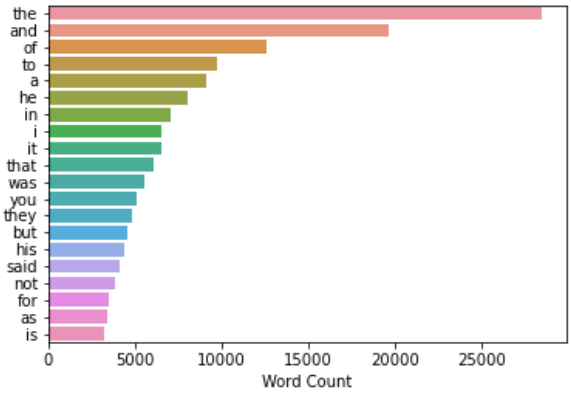

Figure 1: Top 20 Words in the LOTR Corpus

Hence, the text data will require additional cleaning in the early stages of analysis in order to reduce the ratio of noise to useful information and to effectively answer the research questions.

The second data component, the demographic data on the characters, contains 911 rows, each corresponding to a distinct character, as well as 9 columns, record information about their date of birth, date of death, gender, hair color, height, name, race, realm of origin, and name of their spouse.

However, the data is characterized by incompleteness. This is shown by figure 2, which plots the data frame and marks every cell with a missing value in black.

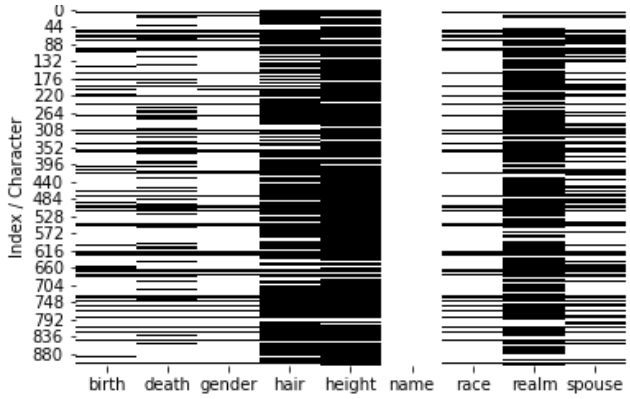

Figure 2: Missing Values of Character Demographics Data

Counting the number of times each name is mentioned in the trilogy revealed that 729 characters never appear in the LOTR books. Further investigating revealed that the data also includes characters in other books about middle earth like 'The Hobbit' or the 12 volume book series 'History of Middle-earth'. In conclusion, to analyze the character dynamics in the LOTR books in a meaningful way, the characters must be filtered down to only include the most important ones.

4 Analysis

4.1 Approach

This section explains the general analytical approach that is used in the subsequent chapters and goes more in-depth on it is applied to answer the research questions of this study. Figure 3 shows a graphical representation of the approach. It is divided into three distinct steps: Data Preparation, Analytics, and Human Reasoning.

The data preparation step aims to convert the data into a format that facilitates further processing and functions as the foundation for all the subsequent analyses. The analytics step then uses more complex data derivation and modeling techniques to uncover latent patterns in the data. Here, visualization bridges the gap between the purely computational methods, and the last step in the analytical approach, human reasoning. Finally, human reasoning uses the capabilities of human cognition, pattern recognition, and domain knowledge to derive conclusions about the data and the analysis methods used. These new insights then allow the analytics step to be refined and more insights to be generated. This iterative process is illustrated by the feedback loop that links the analytics and human reasoning step. To answer each research question, multiple iterations are necessary, before the final conclusions are reached.

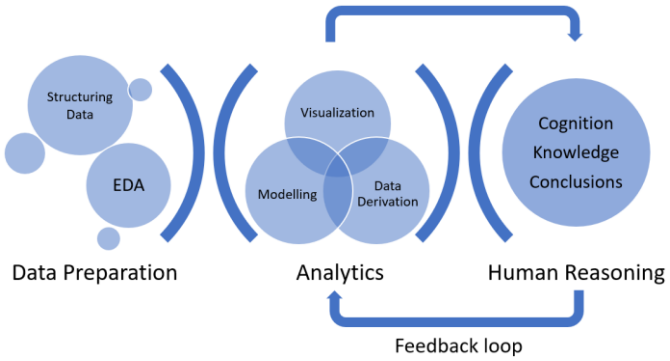

Figure 3: Analytical Approach Illustration

For this study, the data preparation step has a very high significance, given that the majority of data used in this study consists of unstructured text. In order to conduct meaningful exploratory data analysis (EDA) and facilitate further analysis, the text data must be divided into distinct units like e.g., words, sentences, lines, paragraphs, pages, or chapters. This already requires some human reasoning, as the optimal level of aggregation is not obvious and could also vary depending on the use case.

The first research question relates to the performance of word clouds. Here, the role of human reasoning is two-fold. First, it is essential to maximize the potential of word clouds, e.g., by minimizing the amount of uninformative stop words in the text and by using filtering the corpus in a meaningful way to analyze certain parts corpus more in-depth. Secondly, human reasoning is required to recognize patterns in the final word cloud outputs, as this requires an understanding of the meaning and the context of the shown words.

Second, network graphs are an effective tool for visualizing the co-occurrence of keywords and characters and the context in which they appear together. However, human reasoning is required for choosing the hyperparameters that optimize the interpretability of the final graph. Additionally, to derive new knowledge and confirm or deny existing hypotheses about the relationships of the analyzed keywords contextual knowledge is required that only a human analyst can provide.

Similarly, sentiment values and key words occurrences in the story timeline can highlight unique moments and developments in a text. However, to identify overarching patterns and to understand why these highlighted parts might be important, they must be analyzed by a human analyst who has additional information about the context of the story.

In the last step, human reasoning decides when the iterative analytical approach is stopped, and the final results of the analysis can be compiled.

4.2 Process

In this section, the analytical approach described above will be applied to answer the three research questions outlined in section 1.

Data Preparation

The analytical process begins with data preparation in order to mitigate the challenges of analyzing raw text data. Thus, the three parts of the LOTR trilogy were joined and loaded into a pandas data frame. Each cell corresponds to a new line in the text file, which is achieved by splitting the text at the '\n' symbol. However, with a total of 38'443 lines in the text, this representation seemed too granular. Thus, regular expressions were used to find all lines containing a book or chapter heading, and the data was grouped by chapter. The resulting data frame contained 62 rows, corresponding to the 62 chapters, and 4 columns containing the chapter text, as well as the names of the corresponding, part, book, and chapter. Figure 4 visualizes this structure and the word count per chapter.

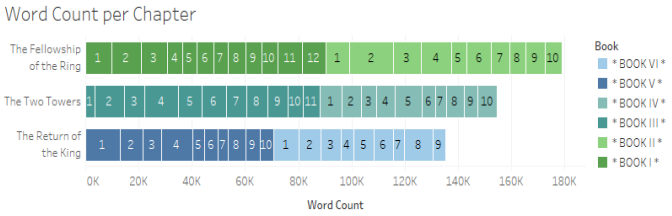

Figure 4: Text Structure and Word Count

This representation already provides insights. It reveals that there are no missing books or chapters, that the chapters are in the correct order, and that they are similar in length. Additionally, the 62 chapters are granular enough to conduct analysis over the story's timeline, while also having the benefit that the content of each text block is summarized by the chapter title.

Question 1: Sentiment and keywords across time.

To make the calculation of sentiment scores more robust, additional preprocessing steps are applied to the text. First, regular expression is used to remove all numbers and punctuation. Next, all words are converted to lower case. A list of stop words from the nltk library is used to remove common words that carry little meaning, as shown in figure 1. Moreover, lemmatization is applied to convert each word to its dictionary form. This is useful as different inflections of a word like e.g., 'saying', 'says' or 'said' are recognized as the same token 'say'. From this cleaned text, a sentiment score is calculated for each chapter. To show deviations from the baseline tone of the story more clearly, the score is normalized by removing the mean from each score and dividing by the standard deviation.

Figure 5 plots the trilogy's normalized sentiment score for each chapter. It can be observed that every two to five chapters, the sentiment shifts from negative to positive and vice versa. Also, the sentiment structure strongly suggests a happy end of the story, as the book ends on six consecutive chapters that have higher than average sentiment. Next, one notices that as the story progresses, it becomes more intense, with the most negative and positive chapters appearing in the last book, Book VI. In book VI, the lowest polarity corresponds to Sam and Frodo reaching Mordor, which is the heart of Sauron's evil, whereas the highest sentiment of the story indicates the chapters where they successfully destroy the ring and peace in middle earth is restored.

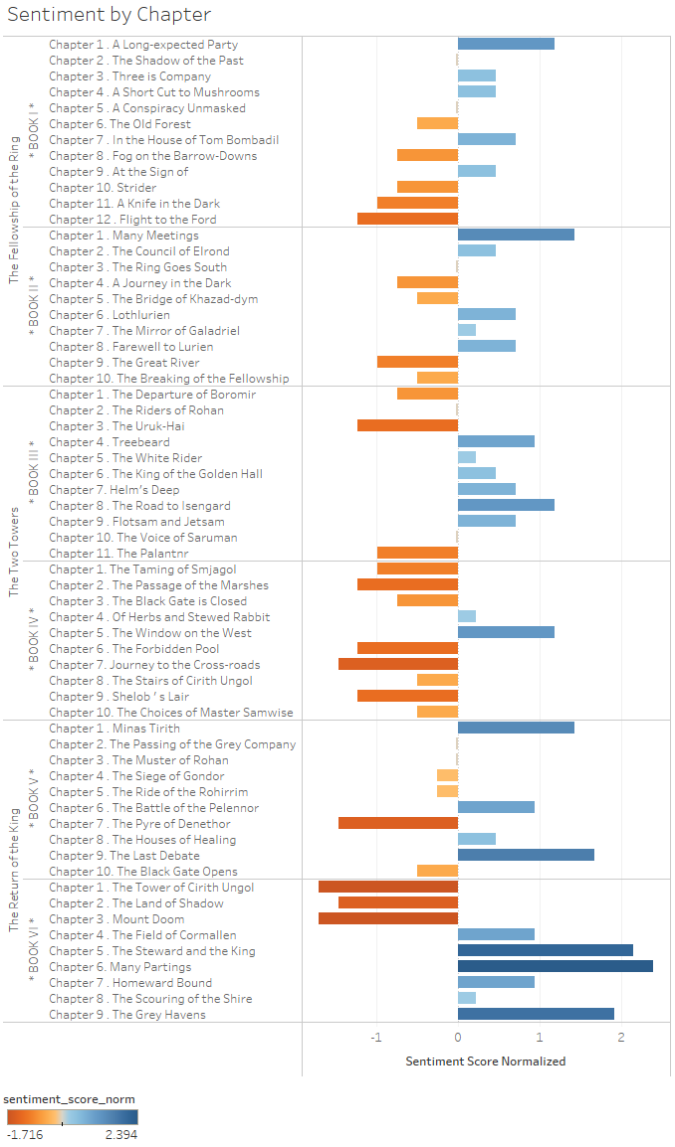

Figure 5: Normalized Sentiment per Chapter

Figure 6 plots a heatmap that visualizes how often a character's name appeared in a chapter. The characters chosen for this analysis consist of all members of the fellowship, as well as Gollum, Sauron, and the ring. A higher number of occurrences is represented by a darker shade of blue. Although the maximum number of mentions was 113, all mentions above 50 are shown with the same color, as this helps to better identify less frequently mentioned characters. Especially the co-occurrences of the characters give a lot of insight into the story. E.g., book I is clearly about the hobbits Frodo, Sam, Pippin, and Merry, as well as the wizard Gandalf. In book II, all characters of the fellowship appear together as they embark on their journey to Mordor. From book III onwards, the heatmap shows how the story alternates between the two parallel storylines of Frodo and Sam on the one hand, and the rest of the fellowship on the other hand, before being reunited in the second half of book VI. Furthermore, figure 6 also helps explain why certain chapters are positive or negative. For example, book IV consists almost exclusively of chapters with below-average sentiment, which coincides with many occurrences of Gollum. Also, the only positive chapter of book IV (chapter 5) is the one in which Gollum does not appear, suggesting that Gollum is closely linked with negative themes in the book.

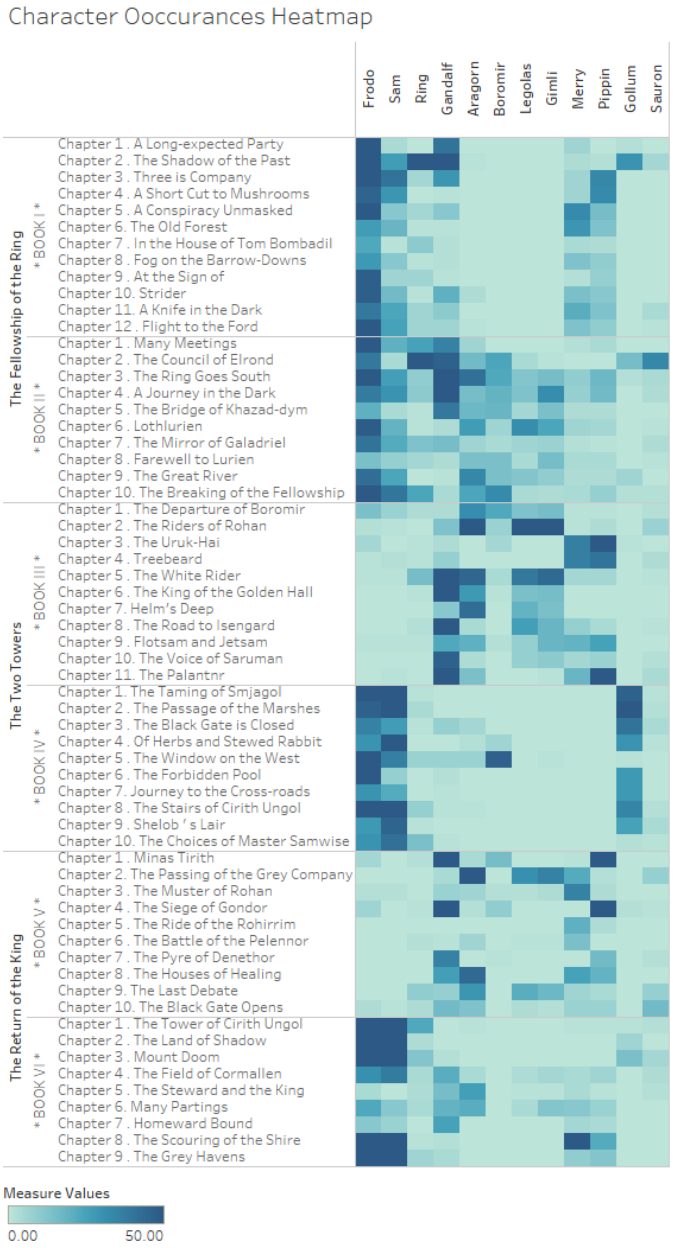

Figure 6: Character Occurrences per Chapter

Question 2: Text summarization with word clouds.

To maximize the interpretability of word clouds, an iterative procedure of filtering, visualizing and human reasoning has to be applied to minimize the share of noise and uninformative words in the text. Thus, similar to the sentiment analysis, frequent stop words are removed from the analysis. To further increase the information density in the text, part-of-speech (POS) tagging is used, which annotates every word in a text with its grammatical role, based on the words definition and its context. This is used to filter the

corpus to only include nouns, as those seemed better suited to capture important concepts and characters in the story.

Figure 7: Word Cloud of LOTR Nouns

Figure 7 visualizes the most frequent nouns in the LOTR trilogy. The word cloud includes all of the fellowship's members, but Gandalf, Frodo, and Sam appear larger than the other six, indicating that they are the most important members. The words day, night, way, and road represent the fellowship's long journey to Mordor to destroy the ring. Also, Sauron appears only infrequently in the word cloud. However, the words hand and eye are very large and refer to ‘the hand of Sauron' and ‘the eye of Sauron', respectively. So, even though his name is rarely mentioned, Sauron seems to be the book's main antagonist.

To further investigate the concept of good and evil, figure 8 shows word clouds that summarize text passages that exceed a sentiment threshold of +0.8 or -0.8. The line-by-line analysis helps to avoid the averaging of sentiments that occurs from analyzing entire chapters, and the threshold further mitigates overlap between the two groups, which increases information density.

Apart from the expected differences in the frequency of words like good and beautiful, versus fear and evil, comparing the size and appearance of keywords sheds light on their role in the story. For example, ‘eye’ and ‘Sauron’ only appear in the negative sentiment cloud, which confirms their negative role in the story. Although the names of the fellowship members appear in both word clouds their larger size in the positive word cloud suggests that they are closer related to the inherently positive things in the story. Lastly, the name Gollum is significantly larger in the negative word cloud, reinforcing the idea of him being connected to negativity and evil. Interestingly, Smjagol – Gollum’s original hobbit name before he got possessed the ring – only appears in the positive sentiment cloud, suggesting that his association with negativity is due to the ring's corruption rather than Gollum's inherent character.

Figure 8: Nouns from High and Low Sentiment Lines

Question 3: Network analysis of most frequent characters.

Network analysis is a used tool to visualize the character interactions in the story. Using the comprehensive data on character demographics, only those characters, whose first name appears over 45 times in the text are included in the final network, resulting in 21 characters. This procedure decreased scope by removing less important characters but in turn, increased the network's interpretability. For the following networks, co-occurrence is defined as two names appearing in the same line.

Figure 9 plots the co-occurrences of the 21 characters. The size of the character node corresponds to the number of connections they have, whereas the color of the line indicates the strength of the connection, measured by the number of occurrences. Frodo has the most connections in the network. This observation, along with his high absolute name frequency, qualifies him as the story’s protagonist. This is also reinforced as he has the highest eigenvector centrality out of all characters, which is a quantitative method of node importance. Moreover, the strongest connections exist among the members of the fellowship, especially Frodo and Sam, the Merry and Pippin, as well as Legolas and Gimli have a strong connection between them. Additionally, Frodo has a strong relationship with his uncle Bilbo and Gollum.

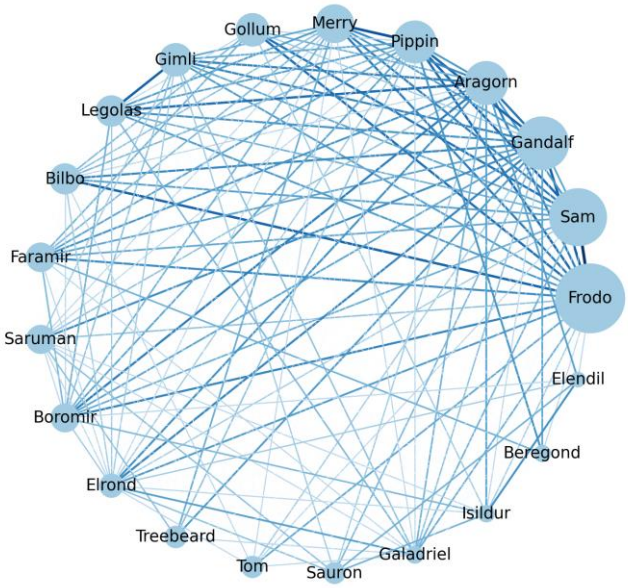

Figure 9: Network Graph of Character Co-occurrences

Figure 10 aims to further investigate the character relationships. By defining the weight of the connection by the sentiment of the line that the cooccurrence they appeared in, the positivity or negativity instead of the strength of the relationship can be analyzed.

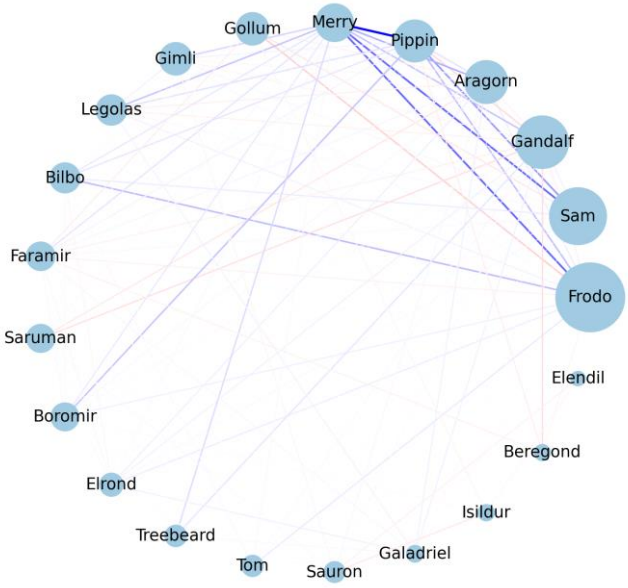

Figure 10: Network with Sentiment Weighted Co-occurrences

The relations between the members of the fellowship are mostly positive. Only Gandalf and Pippin have a negative relations ship, as Gandalf's age and wisdom are directly opposed to Pippin's childishness and ignorance, which leads to some conflict between the two. Also, the rivalry between Gandalf and the wizard Saruman is shown by the negative connection between them. Regarding Frodo's strongest bond with Sam is neural. This seems counterintuitive at first, but it makes sense when you consider that throughout the story, they had to overcome conflicts in their relationship. Frodo's most positive connections are with his uncle Bilbo, and hobbit friends Pippin and Merry. His only strong negative relation is to Gollum, which is in line with the observations about their relationship from previous analyses. It additionally appears to be the strongest negative connection among any character combination in the text.

4.3 Results

Sentiment analysis per chapter revealed deeper insight into the story's structure. It showed that positive and negative emotions get more extreme as the story progresses and that after a final struggle, it concluded with a happy end. Likewise, the character occurrence heatmap gave insight into the parallel storylines after the splitting of the fellowship and indicated that the negative sentiment of book IV is linked to the appearance of Gollum.

Visualizing the complete trilogy with word clouds presented the most important keywords in the story and shed light on the potential protagonists and antagonists of the story. Furthermore, the concept of good and evil was further investigated by filtering the corpus based on sentiment and visualizing lines with high vs low sentiment in two complementary word clouds.

Lastly, network analysis confirms the hypothesis that Frodo is the main character of the story, as he has the most connections with other characters, as well as the highest degree of centrality in the network. Visualizing the average sentiment of the connections reveals a close positive bond between the members of the fellowship in general, as well as conflict between Gandalf and Pippin and the complex relationship between Sam and Frodo.

5 Critical Reflection

The main challenge in this analysis was the unstructured nature of the text data. However, by converting the data into a more structured format, and using computational methods that are designed for text analysis these challenges could largely be mitigated.

Sentiment analysis over the story's timeline proved to be an effective tool, to visualize the flow of the story. However, using a dictionary-based approach to count the number of positive and negative words is subject to certain limitations, that introduce noise to the analysis. For example, this method is not sophisticated to recognize jokes, sarcasm, irony, or deliberate exaggerations. Likewise, this approach has trouble detecting negations and interpreting domain-specific vocabulary. In this study, these limitations are addressed by aggregating sentiment scores on a chapter level, where these imperfections tend to average out and do not influence the analysis significantly. However, using more sophisticated and less noisy methods to assess sentiment, e.g., by predicting sentiment with a language model that also takes context information into account, could be used to conduct a more granular polarity analysis that reveals more nuanced insight into the story's key moments and developments.

The visualization of character occurrences was useful as it added another layer of information to the story's timeline, where patterns could be observed. By analyzing sentiment and character occurrence together, one can develop

hypotheses about links between character and sentiment developments.

Likewise, while word clouds are an effective text summarization tool, their potential is increased when used in combination with other metrics that allow are a more focused investigation of certain concepts in the text. In the study, after studying how chapters and time are related to sentiment, word clouds revealed the most frequent keywords that appear in a positive and negative context. Often the simplicity of keyword-based analyses, like the character occurrence heatmap or the word clouds can be regarded as a strength. However, it also has some disadvantages, especially because only identical words are counted in the frequency measurement. Thus, when a character is referred to in the third or first person, or a nickname it will not be recognized.

Network analysis was able to uncover insights, about the most important characters, the strength, and the polarity of the relations between them. Compared to the other analyses, it also required comparatively little domain knowledge because the characters included were selected based only on frequency and because the strength and sentiment between them can be easily read. This makes network analysis well suited for the application in other texts, as only interpreting the reasons for a certain relationship has to be analyzed in context. Nonetheless, it is affected by the same drawbacks as other analyses that use keywords and sentiment metrics. Additionally, static network analysis limits the number of nodes that can be meaningfully analyzed. Here, an interactive visualization might be able to better represent large and complex networks.

In summary, despite its limitations, the visual analytics applied in this study constituted a practical approach for analyzing the underlying source material and for answering the research questions.

Table of word counts

Problem statement	250
State of the art	500
Properties of the data	500
Analysis: Approach	500
Analysis: Process	1500
Analysis: Results	200
Critical reflection	500

REFERENCES

The list below provides examples of formatting references.

[1] M. Gentzkow, B. Kelly, and M. Taddy (2019). Text as Data. Journal of Economic Literature, 57(3), 535–574. https://doi.org/10.1257/jel.20181020

[2] T. Gu, Lord_of_the_ring_project, (2018), GitHub repository, https://github.com/tianyigu/Lord_of_the_ring_project

[3] S. M. Mohammad (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. Decision Support Systems, 53(4), 730-741.

[4] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl (2014). Word cloud explorer: Text analytics based on word clouds. *47th Hawaii International Conference on System Sciences*, 1833-1842. IEEE.

[5] P. Mutton (2004). Inferring and visualizing social networks on internet relay chat. *Eighth International Conference on Information Visualisation*, 35-43. IEEE.